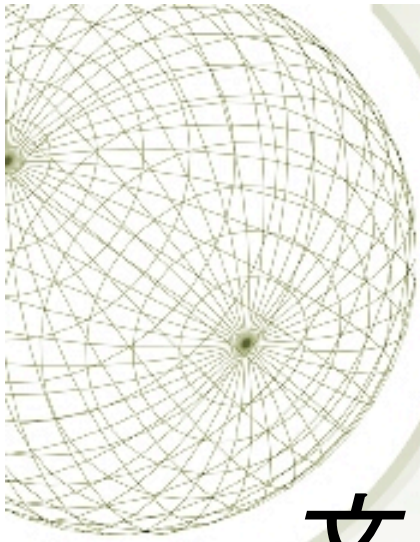
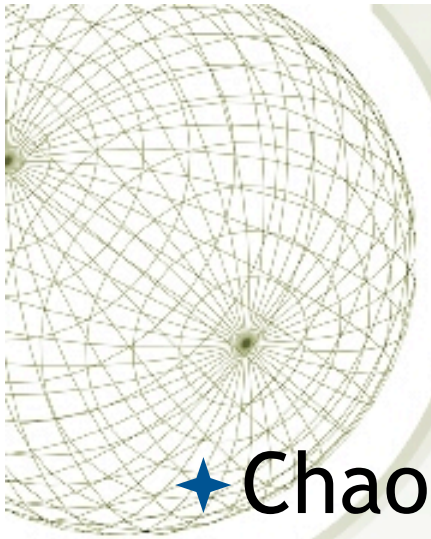


文字オントロジに基づく文字オブジェクト列間の編集距離

師　茂樹（花園大学）



目的

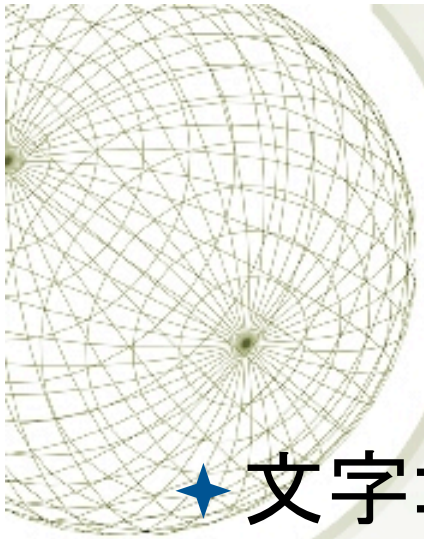
- Chaonモデルの文字オブジェクト間で編集距離を求めたい
  - 文献学への応用（個人的な願望）
    - 写本の比較など
  - 様々な応用

編集距離

- Vladimir Levenshtein氏（1965年）
- 置換・挿入・削除の最小回数（コスト）
  - 例：
    1. 京都大学
    2. 首都大学（「京」を「首」に置換）
    3. 首都大学東（「東」を挿入）
    4. 首都大学東京（「京」を挿入）
- 動的計画法

編集距離の文字コード依存

- 文字コードのモデルの問題
  - 本質主義的文字観
    - 例：Unicodeの*character*
  - 字形中心
- 置換コスト計算の単純さ
  - 有→無のコストと无→無のコストは同じ？
  - 芸（ゲイ）→芸（ウン）の置換コストは0？



問題の所在

- 文字コードに依存しない編集距離
  - 文字コードから文字オブジェクトへ
    - 野村雅昭氏「同字と別字のあいだ」
    - Chaonモデル
  - 文字オブジェクト間の距離

「同字と別字のあいだ」(1)

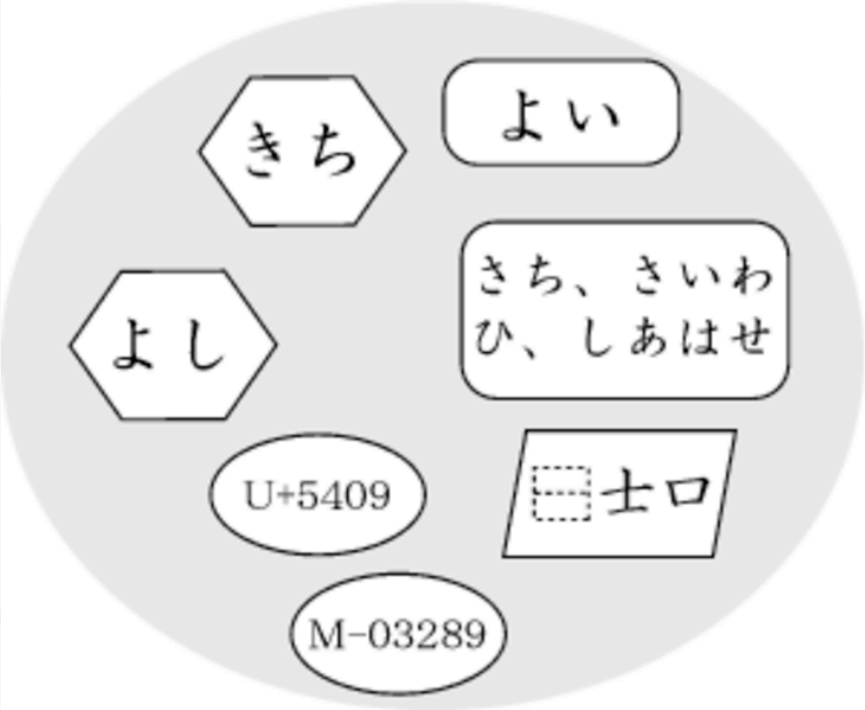
- 野村雅昭氏の文字比較モデル(1984)
- 字体素・音素・意義素による比較
  - 単純すぎる面も？

「同字と別字のあいだ」(2)

	形	音	義	例
①	＝	＝	＝	（同字）
②	＝	＝	≠	（該当例なし）
③	＝	≠	＝	（該当例なし）
④	＝	≠	≠	芸（ゲイ）—芸（ウン）、缶（カン）—缶（フ）
⑤	≠	＝	＝	単—單、歯—齒、円—圓、亀—龜
⑥	≠	＝	≠	知—智、編—篇、付—附、激—劇
⑦	≠	≠	＝	足—脚、暖—温、作—製、使—用
⑧	≠	≠	≠	（別字）

Chaonモデル (1)

- 素性の集合による文字の表現
  - 文字オントロジ

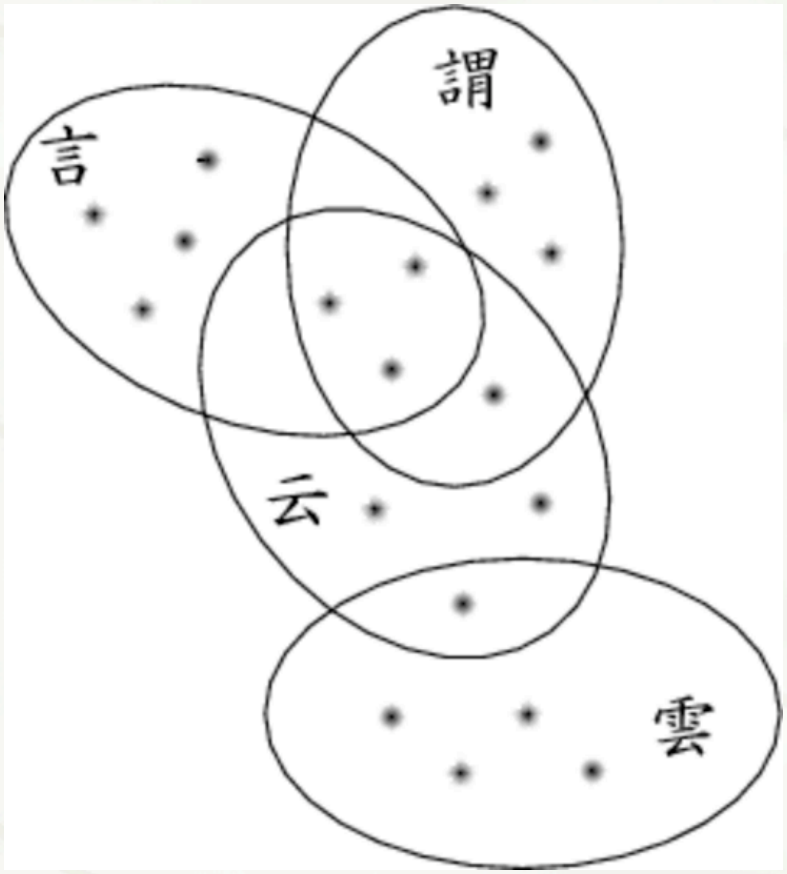


Chaonモデル (2)

- 素性名の階層化
  - 例：=jis-x0208@1997
- 素性値の持つ構造
  - IDS
  - 音韻（子音、母音、声調など）

文字オブジェクト間の距離 (1)

- 集合演算
  - 素性名のマッピングによる比較



文字オブジェクト間の距離 (2)

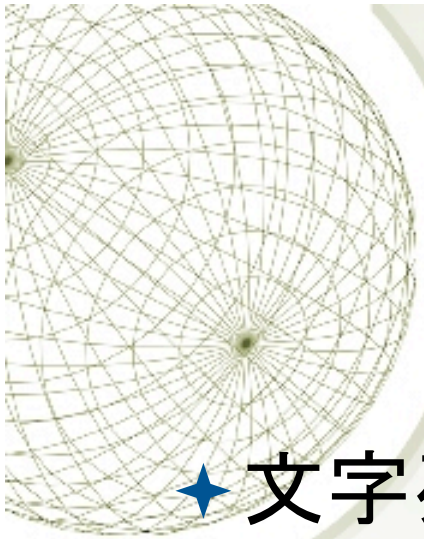
- 素性名のマッピングによる比較

素性名	「雲」	「云」	コスト
形	雨＋云	云	「雨」挿入（コスト0.5？）
音	ウン	ウン	（コスト0）
義	くも	いう	置換（コスト1）

文字オブジェクト間の距離 (3)

- 素性名が階層化されている場合 (1)
  - 単純な比較

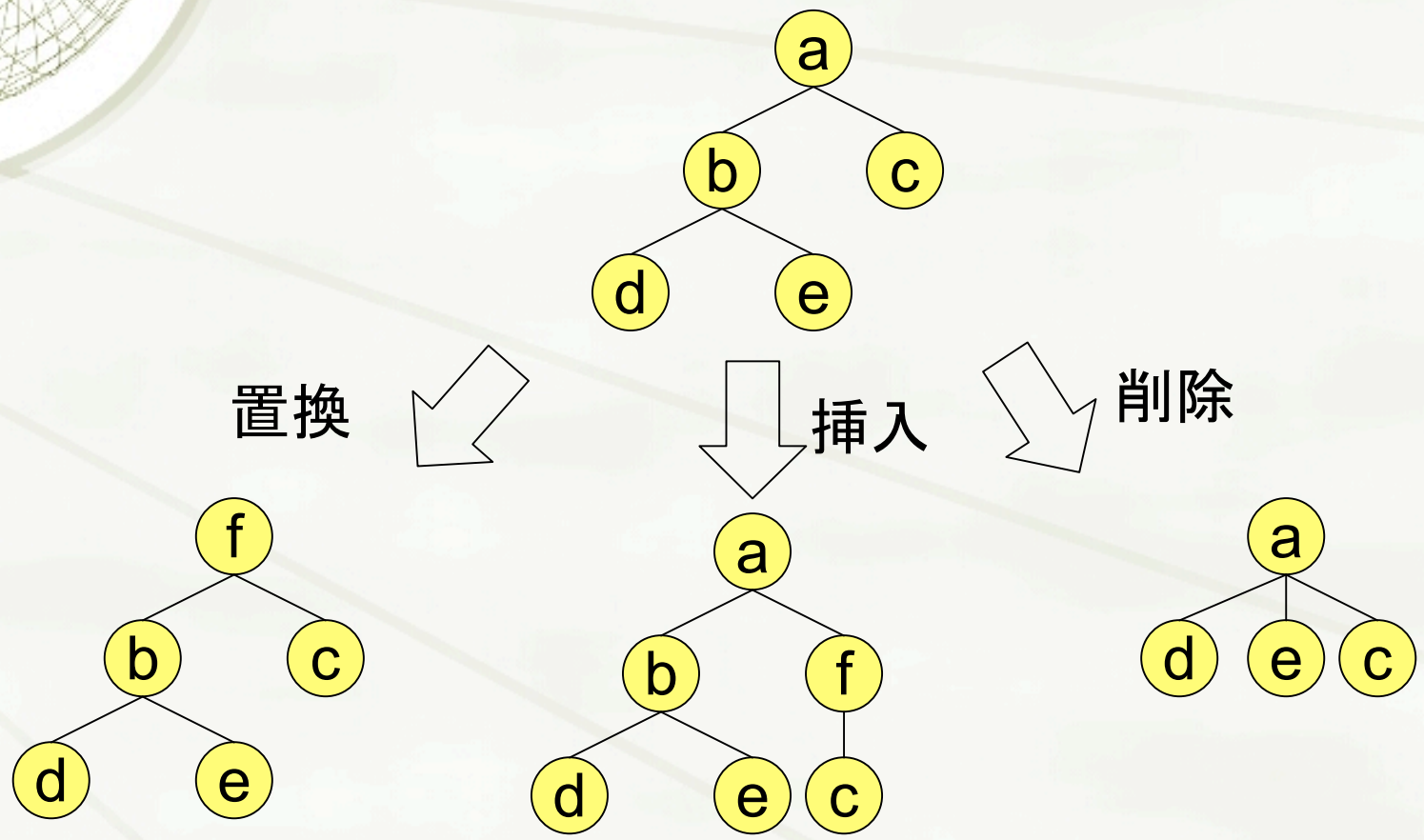
素性名	呉A	呉B	コスト
jis-x0208@1997	3862	×	削除（コスト1）
jis-x0208	×	3862	追加（コスト1）



木の編集距離 (1)

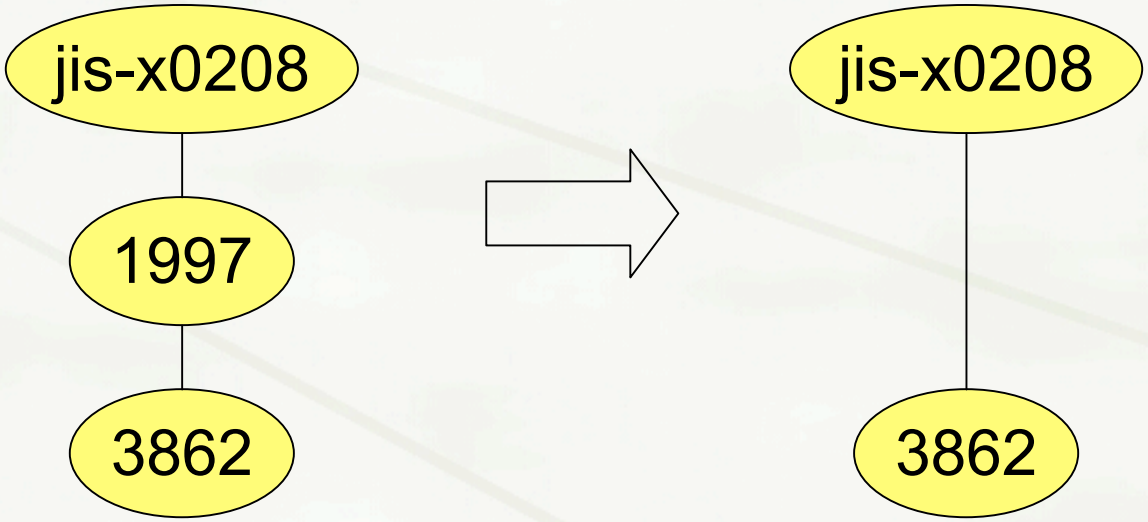
- 文字列の編集距離を拡張
  - 多くの研究
- 置換・挿入・削除の最小回数(コスト)

木の編集距離 (2)



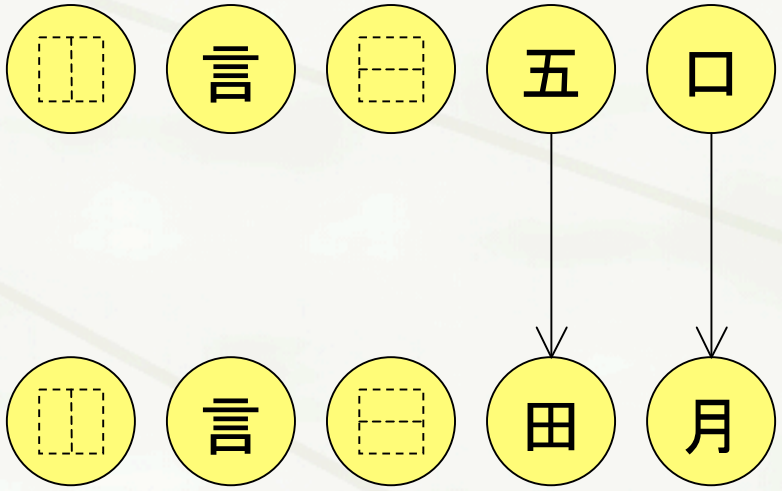
文字オブジェクト間の距離 (4)

- 素性名が階層化されている場合 (2)
  - 木の編集距離として



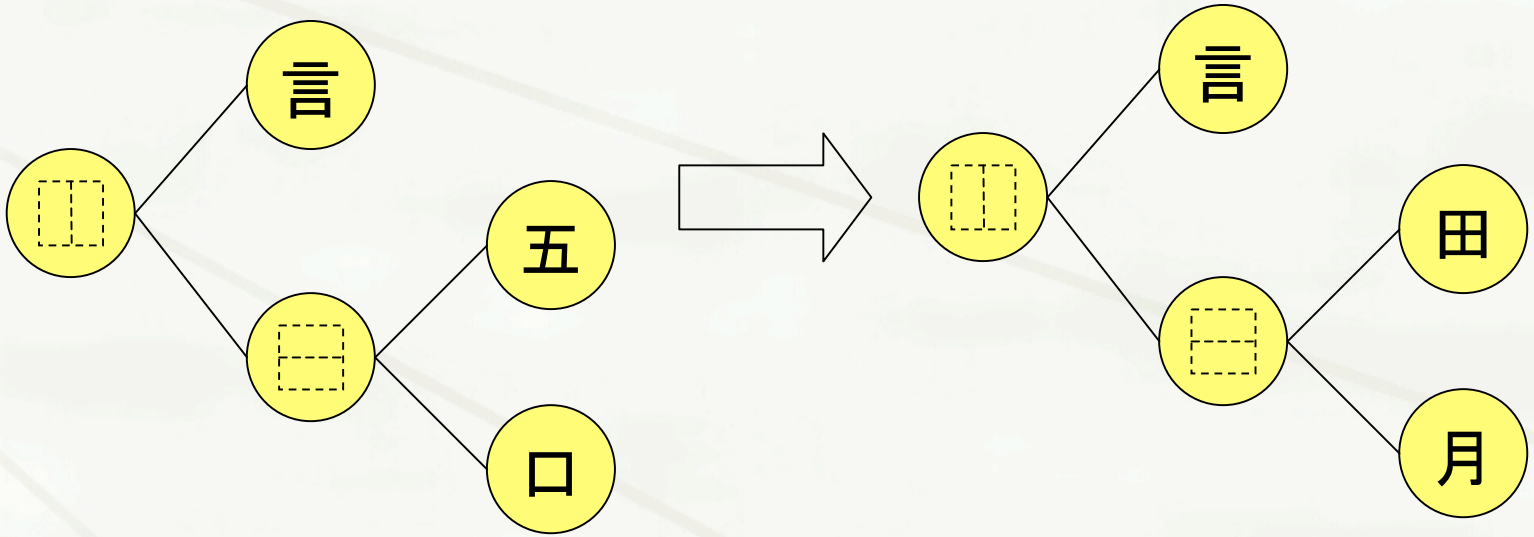
文字オブジェクト間の距離 (5)

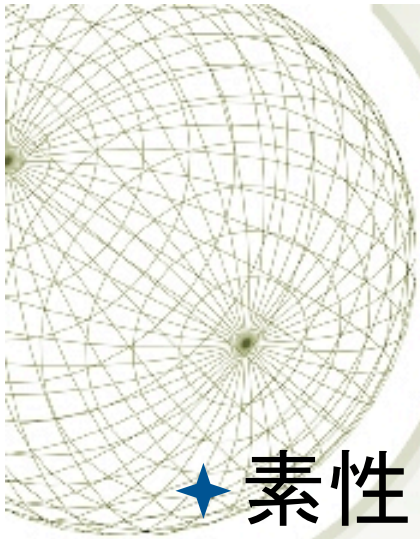
- IDSの編集距離 (1)
  - 文字列の編集距離として処理



文字オブジェクト間の距離 (6)

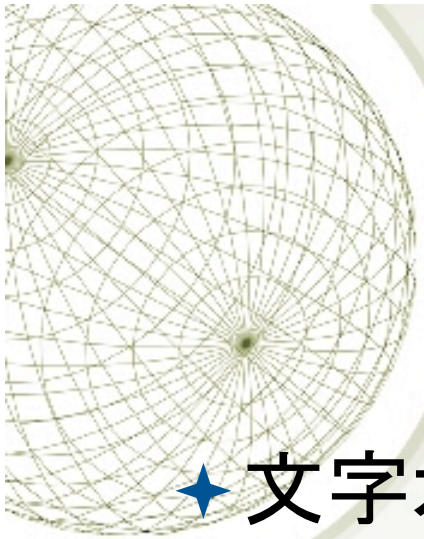
- IDSの編集距離 (2)
  - 木の編集距離として





木構造にするメリット

- 素性による処理の場合分けをしなくてもよい（かもしれない）
- 文字列も木構造の集合（森）として考えられる

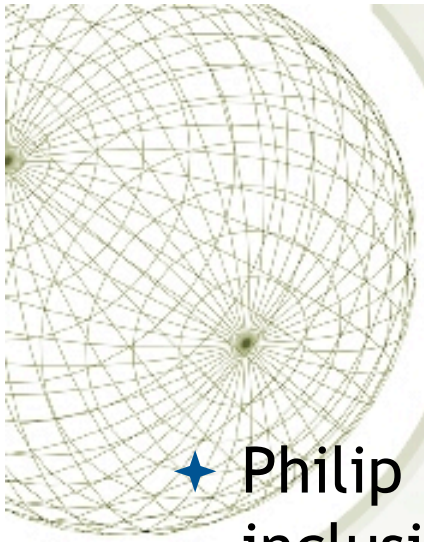


問題点

- 文字オブジェクト木の無限後退
- 各種構造の正規化
- データベースの充実
  - 少なくとも形・音・義は揃わなければ
- 様々なコスト
  - 計算量
  - 面倒くさい

文字オブジェクト列間の距離

- 文字オブジェクト木の順序付き集合（森）間の編集距離

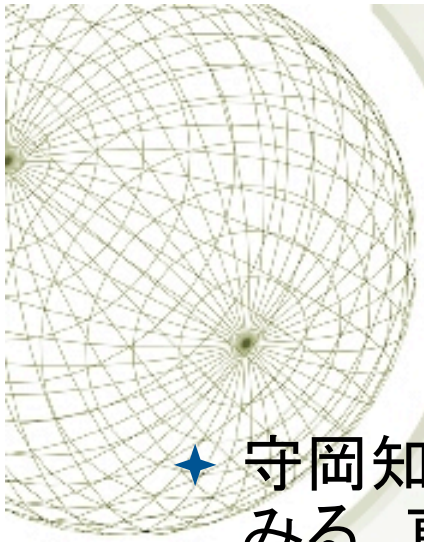


参考文献 (1)

- Philip Bille. Tree edit, alignment distance and inclusion. Technical report *TR-2003-23* in *IT University Technical Report Series*, Mar 2003.
- Kuo-Chung Tai. The tree-to-tree correction problem. *Journal of the Association for Computing Machinery*, Vol. 26, 1979.
- Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, Vol. 18, No. 6, 1989.

参考文献 (2)

- 秋山陽一郎, 守岡知彦, 浦田衣里. 階層的素性名を用いた異体字記述の試み. 情報処理学会研究報告, Vol. 2005, No. 76, pp. 55-61, Jul 2005. 人文科学とコンピュータ研究報告2005-CH-67.
- 久保山哲二, 宮原哲浩. 木の編集距離を用いた半構造データからの情報抽出. 第18 回人工知能学会全国大会講演論文集, 2004.
- 野村雅昭. 同字と別字のあいだ. 日本語学, Vol. 3, No. 3, 1984.
- 守岡知彦, 師茂樹. 文字素性に基づく文字処理. 情報処理学会研報告, Vol. 2004, No. 58 (2004-CH-62), May 2004.



参考文献 (3)

- 守岡知彦. CHISE で複数の文字同定規準をサポートしてみる. 東洋学へのコンピュータ利用第16 回研究セミナー, Mar 2005.
- 師茂樹. Perl/CHISE による正規表現の拡張の試み—文字素性による後方参照の実装実験と課題—. Linux Conference 抄録集, Vol. 1, 2003.
- 師茂樹. N グラムと文字データベースによる漢字仏教文献の分析. 情報処理学会研報告, Vol. 2004, No. 7, Jan 2004 (2004-CH-61).
- 師茂樹. Surface or Essence: Beyond the Coded Character Set Model.「書体・組版ワークショップ」報告書, Feb 2004.

参考文献 (4)

- 師茂樹. Unicode の*character* 概念に関する一考察. 東洋学へのコンピュータ利用第15 回研究セミナー, Mar 2004.
- 師茂樹. 思想史としての文字情報処理: 問題提起として. シンポジウム「文字情報処理のフロンティア: 過去・現在・未来」予稿集. 花園大学国際禅学研究所漢字処理研究室, June 2004.
- 矢野環. 芸道伝書の発展経過の数理文献学的考察 —Spectronet, Split decomposition—. 情報処理学会研究報告, Vol. 2005, No. 10 (2005-CH-65), 2005.